Structural Analysis of fMRI Data Revisited: Improving the Sensitivity and Reliability of fMRI Group Studies

Abstract

Group studies of functional MRI datasets are usually based on the computation of the mean signal across subjects at each voxel (Random Effects Analyses), assuming that all subjects have been set in the same anatomical space (normalization). Although this approach allows for a correct specificity (rate of false detections), it is not very efficient, for three reasons: *i)* its underlying hypotheses, perfect coregistration of the individual datasets and normality of the measured signal at the group level, are frequently violated ; *ii)* the group size is small in general, so that asymptotic approximations on the parameters distributions do not hold ; *iii)* the large size of the images requires some conservative strategies to control the false detection rate, at the risk of increasing the number of false negatives. Given that it is still very challenging to build generative or parametric models of inter-subject variability, we rely on a rule based, bottom-up approach: we present a set of procedures that detect structures of interest from each subject's data, then search for correspondences across subjects and outline the most reproducible activation regions in the group studied. This framework enables a strict control on the number of false detections. It is shown here that this analysis demonstrates increased validity and improves both the sensitivity and reliability of group analyses compared with standard methods. Moreover, it directly provides information on the spatial position correspondence or variability of the activated regions across subjects, which is difficult to obtain in standard voxel-based analyses.

Index Terms

functional MRI, Group analysis, spatial normalization, structural methods, watershed, belief propagation, replicator dynamics, group comparison.

Structural Analysis of fMRI Data Revisited: Improving the Sensitivity and Reliability of fMRI Group Studies

I. Introduction

Functional neuroimaging aims at finding brain regions specifically involved in the performance of cognitive tasks. In particular, functional MRI (fMRI) is based on the detection of task-related Blood Oxygen-Level Dependent (BOLD) effect in the brain. The measurement of this effect is performed by regression analysis of four-dimensional datasets (three spatial dimensions plus time) against pre-defined regressors that represent the expected BOLD response to the stimulations across time; this analysis framework is known as the General Linear Model (GLM) [1]. Inference about putative regions of activity is generally based on several subjects ($\sim$10-15 subjects typically), and the current standard procedure consists in detecting voxels for which the average task-related BOLD signal increase is significant across subjects (random/mixed effects analyses, R/MFX) [1], [2]. Such voxel-based inference schemes require the images to be warped to a common space, which is usually performed by coregistration of the anatomical, then functional data with a template image [3]. In most data analysis software packages, the reference image is the average T1 image provided by the Montreal Neurological Institute (MNI), which matches approximately the Talairach coordinate system [4].

Voxel-based inference schemes are explicitly based on the assumptions that *i)* the functional images are properly co-registered, so that a location in the common space corresponds to the same region in the brain of each subject; *ii)* at a given spatial location in the reference space, the signal is normally distributed across subjects, so that the RFX and MFX statistics are Student-distributed under the null hypothesis that no activation occurs. Both hypotheses might be wrong: the signal can be inhomogeneous across subjects [5], so that normality assumptions are not met [6], and mis-registrations remain after spatial normalization of the datasets. The magnitude of such local shifts is probably 1cm in many brain regions (this can be observed for functional regions like the the motor cortex or the visual areas [7], [8] or the position of anatomical landmarks [9]–[11]). In addition, the number of subjects included in the analysis is generally small, so that RFX analyses are known to have a weak sensitivity.

In order to deal with the spatial mis-registration issues, most neuroscientists are thus accustomed to smoothing their datasets (8-12mm FWHM typically in group studies) to increase signal spatial overlap across subjects. This leads to biased and less precise localization of activated regions and may in some cases reduce sensitivity. The interpretation of the boundaries of supra-threshold

regions in group studies is not clear. Another approach consists in computing local or global anatomical warps that improve inter-subject co-registration [12]–[14]. But such warps may require the additional use of anatomical landmarks, and it is not clear that different brains can be correctly warped onto each other. In particular, the variability in the large scale sulco-gyral anatomy [15], [16] might imply that no such correspondences exist. Note also that Talairach atlas was designed for sub-cortical structures.

In order to cope with non-normality of the signal across subjects, robust inference schemes, based e.g. on the sign test or Wilcoxon signed rank's statistic [17] have been designed. Moreover, permutation-based assessment of the group signal statistics [18], [19] yields an unbiased significance for the statistical maps across subjects, and thus bypasses some approximations implied by the use of random field theory [1].

However, performing a test on each and every voxel has a statistical cost (multiple comparison correction of the p-values), while many of these voxels are probably of little relevance to the cognitive function under study. An interesting alternative is thus to perform inference at a higher level than the voxel level. In other words, one can consider functional regions, or structures, that are found active across subjects rather than active voxels. This point of view has been advocated by many groups that would use functional localizer paradigms to define brain regions before testing the activity of these regions in other conditions [20]. In particular, regions of interest are frequently defined anatomically in order to ease functional studies [21]–[25]. However, such regions are defined within a reference space (e.g. MNI space), which raises the aforementioned issue of mis-registrations; moreover, such approaches define regions very coarsely [21], [25] (less than hundred regions for the entire brain). It is thus necessary to propose data-driven approaches.

In the literature, there is no generally accepted generative model of brain activity that could drive group inference procedures. Although few attempts have been proposed recently [26]–[28], such approaches are likely to be confounded by the complexity of the data, the unknown extent and nature of the activations networks and the global cross-subjects variability. Therefore, a more pragmatic solution consists in modelling some structures of interest observed in the groups of subjects, and then to compare them in order to infer a group-level template of the observed data. Such approach are rule-based rather than based on a generative model of the data. Hereafter, such approaches will be called *structural*.

Structural approaches have to address several important questions:

- What are the structures of interest in each subject? In the case of fMRI data, it is clear that the information of interest is coded in the maxima of activity maps, e.g. large supra-threshold clusters [18], [29], scale-space blobs [30] or activity peaks [31]. Alternatively, some alternative approaches start with the prior definition of regions (parcels), based on clustering of anatomical and/or functional datasets [7]. Some of these approaches might be somewhat coarse for a fine description of activated areas [7], [29], [31]. In this work, we rely on watersheds of supra-threshold areas, which is an intuitive and classical technique in pattern recognition [32].
- How to associate such regions across subjects ? This point may be more difficult, in particular because there exists clearly no isomorphism between individual active regions. While the position in a common space is an important information [29], [31],

there might be local variations that induce some ambiguities. In such cases, the relative position of neighboring regions might be of great importance [7], [30]. In this work, we propose a relatively simple scheme to take this information into account.

- How to validate the sets of regions that have been associated across subjects ? In [30] a procedure that takes into account the individual feature quality, structural similarity between features and association strength, has been proposed. However, its complexity may be quite problematic for interpretation purposes. Here, we prefer to perform a spatial density test on the candidate regions, which allows a strict control on the specificity (type I error rate) of the method.

Finally, another important point is that structural methods involve many parameters in the modelling steps, and it is thus quite important to control the robustness of the results with respect to mild variations in the parameter setting.

In the present paper, we propose a framework that solves the aforementioned issues sequentially; in brief 1) it extracts regions of interest (ROIs) in each subject's dataset, 2) tests which of these regions are reasonably close to other activated regions in other subject's datasets, 3) searches for probabilistic correspondences of the regions across subjects so that the relative positions of ROIs coincide, 4) builds clusters of inter-subject corresponding regions. Such clusters will be termed *cliques* in this paper. Group inference proceeds through the definition of spatial confidence regions associated with each clique, while each subject may or may not have a region associated with a clique defined at the group level. Thus, the method results consist in a group-level model *and* individual instances of this model. This gives some means to account for and characterize inter-subject differences, a key issue in group studies [6], [33].

We describe the method in Section II, and some artificial and real benchmark datasets in Section III. Importantly, our approach allows for an explicit control on specificity, which is shown in Section IV; in Section V we illustrate the improvement in terms of sensitivity and reliability of fMRI group analyses. Reliability is assessed by jackknife subsampling in a population of 102 subjects, and we show that the results of the proposed method are less dependent on the particular subgroup of subjects under study than standard voxel-based tests. Finally, we describe the results of the method when applied to the whole group of 102 subjects. Technical issues and implications for neuroimaging studies are discussed in Section VI.

II. Methods

A. Notations

Let us assume that a group of S subjects take part in an fMRI acquisition protocol while they undergo a certain cognitive experiment. After some standard preprocessing (distortion correction, correction of differences in slice timing, motion correction, normalization), the dataset of each subject is analysed in the General Linear Model (GLM) framework: for a given subject $s \in \{1, .., S\}$, let Y^s be the dataset written as matrix (scans$\times$voxel), and let X be the design matrix that describes effects of interest and confounds; the GLM proceeds by estimating the effect vectors β^s such that

$$Y^s = X\beta^s + \epsilon^s, \ \forall s \in \{1, .., S\}, \qquad (1)$$

where ϵ^s represents the residual matrix. The estimation is based on a maximum likelihood approach performed in each voxel, where the noise is assumed to be an AR(1) process [1], [2], [34]. Let c be the linear combination of

the experimental conditions that is of particular interest; c is also called a functional contrast. A certain statistic ϕ^s can be computed in each subject s to assess the presence of a positive effect $c^T\beta^s > 0$ in each voxel of the dataset, e.g.

$$\phi^s(v) = \frac{\mathbb{E}(c^T\beta^s(v)|Y^s)}{\sqrt{var(c^T\beta^s(v)|Y^s)}} \quad (2)$$

at each voxel v.

Our method takes as input the activations maps ϕ^s of each subject $s \in \{1, .., S\}$, which can be thresholded at a certain significance level, using either voxel-level or cluster-level assessment. In what follows, we assume that, given the significance level $\mathcal{P}$, there exists a known threshold θ_0 such that $P(\phi^s(v) > \theta_0|H_0) < \mathcal{P}$ at any voxel v, where H_0 represents the null hypothesis that no activation is present. Our analysis procedure is illustrated in Fig. 1 and consists of four steps, which are detailed in the next parts.

[Figure 1 about here.]

B. Intra-subject Structural analysis

In the absence of a sound prior on the nature or the position of activated foci, our approach first extracts regions of interest in each dataset. It seems particularly meaningful to segment the main peaks of activity within the supra-threshold components of the statistical maps $(\phi^s)_{s\in\{1,..,S\}}$: the connected supra-threshold components in each subject s are thus segmented into $I(s)$ regions using a watershed method, so that each segmented region is associated with a local maximum of the map ϕ^s. Let $(a_i^s)_{i=1..I(s)}$ be the corresponding maxima for subject s, and (t_i^s) their MNI coordinates (which approximate Talairach coordinates). It is useful to have a graphical representation of the spatial relationships between the segmented regions in each subject (see Sec. II-D). Several models may be used to build such a graph, for instance the neighboring relationships between adjacent regions. The resulting graphs $(\mathcal{G}^s), s \in \{1, .., S\}$ are undirected, and they may contain cycles. We propose an alternative, that produces acyclic graphs: The list of maxima of any connected regions can be organized according to the order relation $\mathcal{O}$: $a_i^s \mathcal{O} a_j^s$ if and only if the corresponding regions are neighboring and if a_j^s is the highest maximum in the vicinity of a_i^s. The set of these structures across regions defines a directed acyclic, possibly disconnected graph in each subject. We consider the *undirected* graphs with the same edges; these will be denoted $G^s, s \in \{1, .., S\}$. Assuming that the activated regions are aligned along some sulci, hence in one-dimensional structures, the tree-like representation given by G^s may code quite well their spatial organization.

Note that the normalization procedure can, and in principle should, take place *after* the intra-subject analysis part. We have implemented both solutions on a real dataset, and have not noticed any significant difference in the global outcome of the method.

C. Spatial statistics

Given that the initial threshold $\mathcal{P}$ should preferably be kept low to avoid false negatives, the first step necessarily results in several false positives. A statistical test on the spatial distribution of the maxima is thus performed to control the false positive rate. Only regions with across-subject reproducible activity are of interest. Thus we build a spatial statistic to remove the local maxima of each subject that are far from local maxima of other subjects. This spatial statistic is the density of presence of supra-threshold local maxima in the other subjects. Let $\tau = (x, y, z)$ be a position in the common

space. For each subject $s \in \{1, .., S\}$, we define

$$\mathcal{D}_s(\tau) = \sum_{\sigma \in \{1,..,S\}-\{s\}} \sum_{i=1}^{I(\sigma)} \exp\left(-\frac{\|\tau - t_i^\sigma\|^2}{2\delta_\tau^2}\right) \quad (3)$$

The parameter δ_τ represents an inter-subject spatial variability and is set to 10mm.

The distribution of the quantity $\mathcal{D}_s(\tau)$ can then be compared, in every location, to its distribution under the null hypothesis H_0. The null hypothesis that we consider here is global, i.e. it means that there is no specifically task-related region in the brain. Under this assumption, the spatial density of local maxima is uniform in the brain volume. We estimate the distribution of $\mathcal{D}_s(\tau)$ under H_0 by random resampling of the position of the activation maxima $(a_i^\sigma)_{i=1..I(\sigma),\sigma \neq s}$ within the brain volume. Let $\tilde{\mathcal{D}}_s$ be the surrogate distribution obtained after k resamplings ($k = 10$ typically).

Then, let α be a significance level, and let u_α be a threshold on the values of $\mathcal{D}_s$ such that $P(\mathcal{D}_s(\tau) > u_\alpha | H_0) < \alpha$; an estimator of u_α is given by the α-quantile of the density $\tilde{\mathcal{D}}_s$

$$u_\alpha = \text{arginf}_u \left[\frac{1}{\Omega}\int \mathbb{I}_{\tilde{\mathcal{D}}_s > u}(\tau) d\tau < \alpha\right] \quad (4)$$

where $\Omega = \int d\tau$ is the brain volume.

Importantly, the test is performed for a small number of spatial locations $(t_i^s)_{i=1..I(s)}$. Hence its significance can be corrected using a Bonferroni procedure, i.e., by replacing α by $\frac{\alpha}{I(s)}$ in Eq. (4). An example is provided in Fig. 2.

One might be concerned with the behaviour of the method, assuming that the null hypothesis has been rejected in some regions of the brain: Does the test remain valid in the other regions, given that the global null hypothesis of a uniform density of maxima has been rejected ? In fact, in such case, the resampled distribution $\tilde{\mathcal{D}}_s$ is an *overestimation* of the true null distribution under the null hypothesis $P(\mathcal{D}_s|H_0)$, which means that the ensuing test is conservative, hence valid. This fact is evident in Fig. 2, where the (null) mode of the resampled distribution is shifted to the right, with respect to the mode of the non-resampled distribution.

[Figure 2 about here.]

The test is iterated in all the subjects, then non-significant maxima at the desired significance level are rejected. The process can be iterated in order to refine the spatial model. Let $\mathcal{I}(s) \leq I(s), s \in \{1, .., S\}$ be the number of remaining regions in each subject. Since we control the probability that one false positive region might show up in a given subject at level α, given $\nu \in \{1..S\}$, the probability of one false positive region in ν subjects over S is controlled by the binomial law $\mathcal{B}(\nu, S, \alpha)$.

D. Finding correspondences using Belief Propagation networks

The statistical procedure leaves us with a set of candidate regions which are spatially clustered across subjects. Then the core part of the procedure consists in finding which regions correspond across subjects. This problem is probably the most difficult one, since one would like to obtain explicit correspondences between individually segmented regions, although there cannot be a one-to-one correspondence across subjects. For instance, a couple of neighboring active foci in one subject might not be distinguishable in some other subject. Our solution consists in estimating, for each pair (s_1, s_2) of subjects, the probability that the region $a_j^{s_1}$ in subject s_1 is the analogue of region $a_i^{s_2}$ in subject s_2. Then, given these probabilities, cliques of cross-subjects regions will be constructed; this will be detailed in section II-E.

We search for inter-subject correspondences in the *relative positions* of activated areas. While this does

not rely on a physical model, it builds on the heuristic that across subjects, positions of activated regions should be locally similar, though not identical in the common space. For instance, it is logical to favor configurations in which couples of neighboring regions with similar relative positions in two subjects will be grouped in two different cliques according to their relative position.

We base our search on a graphical model of the position of the maxima in each subject. This model is provided either by the undirected acyclic graph G^s or the undirected and possibly cyclic graph $\mathcal{G}^s$ defined in section II-B, from which non-significant nodes have been removed. Probabilistic associations are then searched for each pair of subjects, using a belief propagation (BP) algorithm [35]. Given a reference subject s_1 and a target subject s_2, the associations are initialized as

$$P(a_i^{s_2} \leftarrow a_j^{s_1}) \propto \exp\left(-\frac{\|t_i^{s_2} - t_j^{s_1}\|^2}{2\delta_\tau^2}\right) \qquad (5)$$

with appropriate normalization, where $P(a_i^{s_2} \leftarrow a_j^{s_1})$ stands for the probability that the maximum $a_i^{s_2}$ in subject s_2 is the homologue of maximum $a_j^{s_1}$ in subject s_1. These probabilities are refined by belief propagation; for each edge (jk) of the graph G^{s_1} or $\mathcal{G}^{s_1}$, messages are sent from $a_j^{s_1}$ to $a_k^{s_1}$ to quantify the probability of association between $a_k^{s_1}$ and $(a_i^{s_2})_{i=1..\mathcal{I}(s_2)}$:

$$m_{jk}(i) \propto \sum_{l=1}^{\mathcal{I}(s_2)} P(a_l^{s_2} \leftarrow a_j^{s_1}) \exp\left(-\frac{\|(t_i^{s_2} - t_l^{s_2}) - (t_k^{s_1} - t_j^{s_1})\|^2}{2\delta_\tau^2}\right) \qquad (6)$$

with appropriate normalization ($\sum_{i=1}^{\mathcal{I}(s_2)} m_{jk}(i) = 1$). Eq. (6) simply means that whenever the positions of local maxima $t_l^{s_2}$ and $t_i^{s_2}$ in subject s_2 and $t_j^{s_1}$ and $t_k^{s_1}$ in subject s_1 form a parallelogram, the configurations are favored in which $a_l^{s_2}$ and $a_j^{s_1}$ on the one hand, $a_i^{s_2}$ and $a_k^{s_1}$ on the other hand, are associated. As shown in Fig. 3, taking into account the *relative* positions of the maxima improves cross-subjects correspondences by compensating global translation effects.

[Figure 3 about here.]

The beliefs $P(a_i^{s_2} \leftarrow a_j^{s_1})$ and messages are then updated and normalized according to the formal laws of BP [35], [36]. Note that the graphs G^s have no loops, so that convergence is straightforward. We have also used loopy belief propagation, based on the graphs $\mathcal{G}^s$, which did not raise any issue concerning the convergence of the correspondence probabilities. Furthermore, the choice of G^s or $\mathcal{G}^s$ was not found to be crucial in the method. In our experiments, we use G^s by default.

The estimation of the probabilities is performed on each pair of subjects in the group. As an important note, all the quantities used here are asymmetric. In particular, the graphs G^{s_1} and G^{s_2} have a priori different structures, so that the probabilities $P(a_i^{s_2} \leftarrow a_j^{s_1})$ and $P(a_j^{s_1} \leftarrow a_i^{s_2})$ might be quite different after convergence. This is particularly obvious in the case of many-to-one correspondences, given that the number of maxima in a given region may vary a lot across subjects. The next step essentially chooses which of these correspondence are meaningful at the population level.

E. Extracting homologous regions

All the probabilities of all pairwise associations between subjects are then arranged in a common belief matrix B. A row of B contains all the probabilities that a maximum a_i^s of a subject is associated with all maxima $(a_j^\sigma)_{j=1..\mathcal{I}(\sigma),\sigma\in\{1,..,S\}}$ of the other subjects; note that the associations with the other maxima $(a_j^s)j = 1..\mathcal{I}(s)$ within the subject s itself are null. We can also interpret B as the adjacency matrix of a probabilistic association graph between regions that are pooled across subjects.

Some of these associations may not be realistic, but in that case the two maxima should not have strong common associations with other maxima. To deal with such cases, we proceed with the extraction of the maximal cliques of the belief matrix, i.e. groups of maxima that have mutually strong associations. In our case, the association probabilities are asymmetric, thus the maximal clique approach requires that association probabilities are *relatively* high bidirectionally, in a sense detailed thereafter.

Many clustering procedures are possible, e.g. hierarchical clustering techniques, using average or maximal linkage heuristics or replicator dynamics (RD). Since the latter procedure is more data-driven (it does not require a prior definition of the number q of clusters to be found), we describe it in more details, but we also suggest to use average-link agglomerative clustering, where the number of desired clusters is $q = \text{mean}_s \mathcal{I}(s)$ the average number of regions per subject. Note that in that case, the association probabilities are symmetrized.

A formal definition of graph-theoretical cliques is given in [37], in the case where the matrix B is symmetric. They are termed the *dominant sets* of the graph, and their definition relies on two conditions: *i)* that the similarity value of each element of the clique should be high enough with respect to the average similarity of the other elements *ii)* that any element outside the clique should have a weaker similarity with the clique than the elements of the clique, where the similarity values are computed from the affinity values between graph neighbors. Finally, it is shown in [37] that it amounts to define a membership vector x on the graph vertices, and then to solve the program

$$\text{maximize } x'Bx \text{ subject to } x \geq 0 \text{ and } x'u = 1 \quad (7)$$

where u is the vector of ones with the same size of x. Finally, still in the case where B is symmetric, this problem can be solved using replicator dynamics equations to B (see e.g. [38], [39]). Replicator dynamics consist in initializing, randomly or not, a positive vector $x^{(0)}$ whose length is equal to the total number of vertices, and then in iterating the update rule

$$x^{(i+1)} = \frac{(Bx^{(i)}).x^{(i)}}{x^{(i)'}Bx^{(i)}} \quad (8)$$

where $*.*$ stands for the element-wise product. After a few iterations, almost all the components of x vanish, and the other ones correspond to a maximal clique of the belief graph. The clique is removed, and the process is repeated until no non-trivial clique is found. Other rules than the replicator dynamics can be used instead [40], but we experienced that Eq. (8) works efficiently. We noticed that this procedure tends to over-segment the graph B, which is natural due to the restrictive definition of the maximal cliques (or *dominant sets*, see above), but this is not a problematic issue, since further merging of cliques remains possible. For instance, the procedure can be iterated based on cliques instead of regions, thus yielding larger cliques.

Finally, all the cliques that contain maxima from at least ν (e.g. $\frac{S}{2}$) subjects over S are retained.

F. Derivation of a group template

This procedure provides us with clusters of activated regions defined across subjects. It does not require that all subjects are represented for a given activated region, and therefore is able to account for some inter-individual differences. A large cluster means that subjects *typically* have an activated region that corresponds to this cluster.

In order to make group maps, we assume that the positions t_s^i of the maxima within each clique are normally distributed, and thus represent the cliques through their 95% confidence regions (CR) in the common (MNI)

space. Note that the normality hypothesis used here is about inter-subject ROI positions in the reference space, and not about inter-subject signals, which is a key difference with standard techniques; this assumption is used only to define the inter-subject activated regions in the common space.

As a matter of interpretation, these CR regions are quite close to reproducibility maps [41], i.e. maps that count the number of times a voxel is declared active across subjects in some group, because each CR is associated with zero or one particular region in each subject. The interpretation is thus that *a local peak of activity for the proposed task is expected to be observed within the area defined by the CR in a proportion* $\frac{\nu}{S}$ *of the population*. This is quite different from fixed-effects analyses, which disregard inter-subject variability, and random or mixed-effects analyses, which yield the probability that the effect observed in any subject of the population will be positive.

G. Parameters and implementation issues

The method requires few prior parameters: the initial threshold of activity maps $\mathcal{P}$ and especially the spatial p-value α are chosen in order to control the number of false positives. We take typically $\mathcal{P} = 0.001$, uncorrected for multiple comparisons, and $\alpha = 0.2$, but α can be tuned to obtain explicit confidence levels depending on ν (see below). The other parameters are the spatial relaxation distance δ_τ, which we set as a typical inter-subject variability magnitude $\delta_\tau = 10\,mm$, and the number ν of subjects required for the final selection of cliques. ν is important since it explicitly controls the reproducibility of a region across subjects. For instance $\nu = \frac{S}{2}$ yields regions that can be expected to be found in half of the subjects. Note that all these parameters $(\mathcal{P}, \alpha, \delta_\tau, \nu)$ might be changed reasonably without creating inconsistencies.

The control of false positive regions is based on the control performed in each subject in section II-C. For instance, if one were controlling the rate of false detections, Eq. (4) provides us with the probability of one false alarm on any subject, so that an upper bound of the probability of forming a clique with regions from ν subjects over S under the null hypothesis is given by the binomial law $\mathcal{B}(\nu, S, \alpha)$: hence, the probability of getting at least one clique of ν subjects, among S is given by

$$p < \sum_{n \geq \nu} \mathcal{B}(n, S, \alpha) \tag{9}$$

The computation time of the method, implemented with a C/Python code based on numpy and nipy environments (http://projects.scipy.org/neuroimaging/ni/), is about one minute for a dataset of 10 subjects. The complexity of the method is roughly quadratic, given that all pairs of subjects are submitted to the algorithm described in Section II-D. However, the main bottleneck with large datasets is the clique extraction procedure (section II-E), and the computation time is largely data-dependent. For a dataset of 100 subjects, the proposed method take approximately 1h on 3GHz Pentium IV PC running Linux.

III. Evaluation Data

A. Synthetic Noise

The algorithm was tested on synthetic datasets with no simulated activations, in order to ensure that the false alarm rate was controlled using equation (9). The noise only datasets are meant to simulate multi-subject activation maps in which no specific region is activated. Masks of the brain volume were extracted from $S = 10$ subjects in the true dataset; the maps were filled with random normally distributed values and slightly smoothed

(FWHM = 1.17 voxels, i.e. 3.5mm at 3 mm resolution) in order to mimic the intrinsic spatial correlation of fMRI data. The signal magnitude was then corrected to have variance 1 in each dataset. The whole procedure was applied to these maps for values of the parameter α ranging from 0.1 to 0.8 in 0.05 steps, with $\delta_\tau = 10mm$, $\nu = 5$ and $\mathcal{P} = 10^{-3}$. We also tried with different values of $\mathcal{P}$, ranging from 0.05 to 10^{-4}. Finally, the number of detected regions in the group of subjects, over $n = 100$ simulations is reported, and compared to the theoretical bound provided by Eq. (9).

B. Artificial Activations in Synthetic Noise

We also applied the procedure on synthetic data with activation added to the correlated noise. Four distant regions are added with some signal in order to model spatially coherent activity in the group of subjects. The size of the activated regions varied from 20 to 50 voxels. According to the simulation, their mean position (center of mass) was jittered with a magnitude of 0, 1.7 or 3.4 voxels standard deviation, which represents 0 to 10.4mm at 3mm resolution, and is a good representation of group variability in true fMRI datasets. The activation magnitudes were chosen to correspond to a mean SNR of either $-10dB$ or $-6dB$ in each dataset, which corresponds to a realistic SNR in fMRI datasets. In order to model inter-subject differences in the SNR, we let this value fluctuate across subjects in the ranges [-18dB -6dB] and [-11 -3dB] respectively. Once again, this kind of fluctuation is a reasonable model of standard inter-subject variability.

The CR maps, as well as the RFX maps computed on these datasets, are submitted to Receiver Operating Characteristic (ROC) Analysis. False positive and negative rates were computed while the α parameter (CR) or the threshold (RFX) is varied, resulting in a (sensitivity, specificity) plot. Results were averaged over $n = 100$ repetitions of the simulation.

C. Real fMRI data

We used an event-related fMRI paradigm that comprised ten experimental conditions. Subjects were presented with a series of stimuli or were engaged in tasks such as passive viewing of horizontal or vertical checkerboards, left or right click after audio or video instruction, computation (subtraction) after video or audio instruction, sentence listening and reading. Events occurred randomly in time (mean inter stimulus interval: 3s), with ten occurrences per event type.

102 right-handed subjects participated in the study. The subjects gave informed consent and the protocol was approved by the local ethics committee. Functional images were acquired on a 3T Bruker scanner using an EPI sequence (TR = $2400ms$, TE = $60ms$, matrix size=64×64, FOV = $24cm \times 24cm$). Each volume consisted of 34 $4mm$-thick axial contiguous slices. A session comprised 130 scans. Anatomical T1 images were acquired on the same scanner, with a spatial resolution of $1 \times 1 \times 1.2$ mm^3. Finally, the cognitive performance of the subjects was controlled using a battery of syntactic and computation tasks.

fMRI data pre-processing consisted of 1) temporal Fourier interpolation to correct for between-slice timing, 2) motion estimation; for all subjects, motion estimates were smaller than 1mm and 1 degree, 3) anatomo-functional image coregistration and spatial normalization of the functional images in the MNI/Talairach space. This pre-processing was performed using the SPM2 software (www.fil.ucl.ac.uk, [1]). In particular, spatial normalization was performed using default parameters (non-rigid, low frequency deformation with 8*8*7 basis functions [3]); the normalized images were checked in all the subjects to prevent any gross mistake in the image

co-registration. A slight smoothing was performed (5mm FWHM). Standard statistical analysis were also carried out with SPM2, using the usual high-pass filtering and AR(1) whitening.

In the study of the group data, we concentrated on a specific cognitive contrast that shows the activation elicited by the computation task, after video or audio instruction, from which the mere sentence reading/listening effect is subtracted. Such a cognitive contrast is assumed to yield areas specifically activated in the computation task.

In parallel, the grey/white matter interface was segmented in each subject using the Brainvisa software (http://brainvisa.info/), and is used for rendering.

D. Assessment of the reproducibility

Controlling the specificity of the analysis is not sufficient to have reliable brain maps; another concern is to control the risk of overfit in small populations that could result in poor generalization of the regions found to other groups of subjects.

We dealt with this concern by performing the above analysis in ten disjoint groups of 10 subjects sampled from a set of 102 subjects present in the database. We computed an inter-group reliability index by analyzing how often a voxel is declared jointly active across groups, using the procedures described in [6], [41], [42]: The reproducibility map that gives the number of times each voxel is declared active is computed, and its histogram is derived; this histogram is modelled by a mixture of two binomial distributions, and the index $\kappa \in [0, 1]$ measures the accordance of the bimodal model with the data, which in turns reflects the coherence of the binary maps given as input to the model. If κ is close to 0, there is a very little agreement on which voxels are active, while there is a very good agreement if κ is close to 1.

The reliability was estimated from 100 different random splits of the group. This was computed for Random effects analysis (RFX), the same RFX analysis after 12mm FWHM smoothing of the data (SRFX), a Mixed Effects (MFX) analysis [2], an RFX analysis thresholded at the cluster level (CRFX), a Parcel-based RFX (PRFX), and our estimate of confidence regions (CR). More precisely, RFX, SRFX, PRFX and MFX maps were thresholded at the uncorrected $p < 0.001$ level; the CRFX map was built by taking the voxels with a signal significant at $p < 0.01$ uncorrected level, then clusters of connected supra-threshold voxels were formed, and further selected if their size was significant, at $p < 0.05$ corrected level [18]. Lastly, the CR maps contain the 95% confidence regions for the presence of maxima in $\nu = 4$ over 10 subjects, with $\alpha = 0.2$. Note that the PRFX procedure is performed as in [7], and parcels are recomputed for each randomization and sample. The parameters were chosen in order to guarantee that the specificity of the different methods is roughly equivalent, and that the parameters correspond to standard choices.

Additionally, we compared the results of the CR extraction procedure with different graphical models, G^s, $\mathcal{G}^s$, or the trivial graph with no link (i.e. without the belief propagation algorithm). For this purpose, we used a functional contrast that showed region involved in the processing of auditory instructions, because this functional contrast elicits many neighboring activation foci (a_i^s) (about ten in average) in each temporal lobe and each subject.

E. Analysis of the group data

We computed the RFX map and extracted Confidence Regions of the areas activated by the computation task across subjects. Then, we computed the average signal

and position of the regions in each subject, whenever they are defined. We performed some data-driven clustering of these profiles, which yields an assessment of the population homogeneity. Then we regressed this data against side information that was obtained form the subjects: in this case, we used the age of the subjects, their sex, and their ability to perform the mental rotation in 3D of an object, measured by the rate of correct response in a psychological test. The regression procedure reads simply:

$$Y_g = X_g \beta_g + \varepsilon \tag{10}$$

where Y_g is a (subjects, voxel/ROIs) data matrix that represents the average ROI-based or the voxel-based activation signal, X_g a group-level design matrix of size (subjects, regressors) that represents the covariates of interest across subjects, β_g the second-level regression parameters of size (regressors, voxels/ROIs) and ε the residual. Then the voxel- or ROI-based significance of β can be assessed using a standard t-test.

Finally, in the case of the ROI analysis, not only the voxel-based average signal, but also the cross-subject ROI position in the common space can be given as input to model (10). In this case, a chi square test can be used to assess the correlation of the regressors in X_g with the ROI positions.

IV. Simulation Results

A. Controlling the False Positive Rate in Synthetic Noise

We have applied the procedure to n=100 synthetic datasets generated as detailed in section III-A, with different values of the parameter α. The detection rate is compared with the theoretical bound given in Eq. (9). The results are reported in Fig. 4, which shows that the control of false positives is conservative, since the rate of detected regions obtained is below the predicted value. In fact, the control is probably too conservative. The only case where the control may be problematic is that for very low values of α, the correct definition of u_α (see Eq. (4)) requires a very accurate estimate of the right tail of $\mathcal{D}_s$ under the null hypothesis, hence many resamplings. In practical cases, we found 10 resamplings to be sufficient.

We have repeated the procedure with different values of the first-level threshold $\mathcal{P}$, from 0.05 to 10^{-4}. Although the number of false positive depended on the particular value of $\mathcal{P}$ that was chosen, it always remained under the theoretical bound (which does not depend on $\mathcal{P}$).

[Figure 4 about here.]

B. ROC Analysis in Synthetic Activation and Noise

We computed the ROC curve on synthetic dataset with embedded activation in four regions. The exact position of each activated region may vary from 0 to 3.5 voxels, while the amplitude of the response is allowed to vary around a mean of -6dB or -10dB. ROC curves, that represent specificity/sensitivity compromise when detection parameters vary, are presented in Figure 5 for RFX analysis in the proposed CR method, in the different situations. Note that we consider only the part of the curve with a specificity control below 0.01, since weaker controls are of no practical interest. The simulations are repeated 100 times.

[Figure 5 about here.]

Except for the situation with no jitter, the CR method clearly outperforms the RFX method. In addition, the CR method is clearly less sensitive to the SNR level than the RFX statistic. In the no jitter case, the CR method performs better at high specificity levels, but reaches a

plateau at a lower level when the specificity control is weaker. The reason is that the CR is not meant to achieve 100% sensitivity at the voxel level, since the extent of the CR corresponds to the inter-subject variability of the areas. By contrast, the RFX statistic actually tests the presence of activation in each voxel, and thus can detect 100% of the activated voxels in the absence of jitter.

V. Results on Real Data

A. Gain of sensitivity in a small group of subjects

First, we performed a group analysis in a small group of 10 randomly chosen subjects within the whole group. A CR map, obtained at a $p < 0.05$ significance level for the computation-specific contrast is presented in Fig. 6, together with an RFX map, thresholded at a $p < 10^{-3}$, uncorrected, or at the cluster level at $p < 0.05$, corrected. When the RFX map is thresholded at the same significance level ($p < 0.05$, corrected for multiple comparisons) at the *voxel* level, no voxel survives the thresholding procedure.

[Figure 6 about here.]

In spite of the strong type 1 error control, the CR map contains 19 significantly active regions. In particular, it clearly shows symmetric parietal regions involved in the computation task, while these regions are not detected with the RFX procedure (voxel- or cluster-level statistics). This is in agreement with the literature [43] (see also Fig. 9).

In order to test the robustness of the method, we have repeated the experiment with unsmoothed/smoothed data (FWHM=10mm), with higher or lower first-level thresholds $\mathcal{P} = 10^{-2}, 10^{-3}, 10^{-4}$, with the spatial normalization before or after the watershed, or with a small spatial jitter of activation images across subjects (one voxel). All these changes had a very weak impact on the resulting CR maps. For instance, to test the robustness to small spatial shifts, we randomly shifted the datasets from zero or one voxel in one direction (x,y or z) with equal probability (1/7). We performed the RFX and CR analysis of the resulting group data and derived the reproducibility index κ (see Sec. III-D) from ten such group maps. Over 100 repetitions, we obtained a mean value of $\kappa = 0.61$ (range $0.58 - 0.64$), which has to be compared with a mean value $\kappa = 0.42$ (range $0.37 - 0.47$) in the case of the RFX map.

Importantly, the CR procedure not only provides a group-level activity map, but-also explicit correspondences between active regions at the subject-level and the group data. This is illustrated in Fig 7.

[Figure 7 about here.]

B. Between-Group Reproducibility

The reliability of the group analysis method, assessed by reproducibility index estimated across voxels for different splits of the populations into groups of 10 subjects is shown in Fig. 8. This shows that the proposed CR method outperforms RFX, MFX and SRFX, and to a lesser extent, PRFX and CRFX. It is important to note that this procedure is based on the active or inactive status of each voxel, and should not be favorable to non voxel-based analyses a priori. Moreover, the present results are not related to the particular choice of thresholds or p-values, and other reproducibility indexes also yield similar effects (not shown).

[Figure 8 about here.]

Moreover, for a contrast that shows regions involved in the processing of auditory instructions, we found that the CR regions were more reproducible when obtained using either the acyclic graph G^s ($\kappa = 0.589 \pm 0.01$)

or the spatial connectivity graph $\mathcal{G}^s$ ($\kappa = 0.597 \pm 0.01$) than a trivial graph without edges (i.e. without the belief propagation algorithm, see Eq. (6), $\kappa = 0.539 \pm 0.01$).

C. Analysis of a large population

The application of the CR methods to the whole group of 102 subjects yielded $q = 45$ regions, with a corrected p-value of 0.05 of making one false detection. The CR map is presented with the RFX map on these same subjects in Fig. 9. As is usual with large sample sizes, the RFX map shows very wide activated areas. This is is the result of blurring process inherent to the inter-subject variability.

[Figure 9 about here.]

We describe the properties of the 45 resulting clusters in Table I. In particular the anatomical labels of the regions are found in [21].

[Table 1 about here.]

Such a result, formulated in terms of regions, readily indicates possible asymmetries in the spatial repartition of activations across subjects: activation in the Inferior Frontal cortex are found in the right hemisphere only, activations in the Supramarginal cortex, the Angular cortex are found in the right hemisphere only and activations in the Precentral regions are more systematic in the right hemisphere (4 regions) than in the left hemisphere (1 region).

Next, we computed the average signal per region per subject, and tried to characterize the population by unsupervised classification techniques. We call the average signal per region for each subject the *profile*. Based on simple Euclidean distance between profiles, we have performed some agglomerative clustering of the population, using an average linkage procedure. The results are shown as a dendrogram of the subject's profiles in Fig. 10. In this case, it clearly shows that the population is mainly divided into one group of 97 subjects, and 5 isolated subjects. One can conclude that the population is rather unimodal, with a few outliers. A closer inspection of the outlier datasets reveals that four of them had no significant activations, and the last one had an odd pattern of activity, probably confounded by motion or another low-level artifact.

[Figure 10 about here.]

Finally, we regressed the voxel-based activity maps as well as the *profiles* against three regressors of interest defined in each subject: age, sex, and ability to perform a 3D mental rotation (see Eq. 10). We found no effect of age in either case[1].

Concerning sex, we found in the voxel-based analysis a region where the magnitude of the activity is larger for males than for females [$z = 5.35, p < 0.05,$ corrected at $(18, -68, 60)$mm]. As this place is on the posterior edge of the parietal lobe, and not in a significantly activated region, the interpretation of this result is quite unclear. The ROI-based analysis revealed that there was indeed a positive effect on one ROI [$z = 2.73, p < 0.01,$ uncorrected at $(15, -68, 52)$mm], but moreover that there was a significant effect of the sex on the ROI coordinates in MNI space across subjects [$\chi^2_3 = 13.95, p < 0.01,$ uncorrected at $(15, -68, 52)$mm], indicating that there might be some systematic shift effect between males and females. Importantly, no such effect can be observed using the voxel-based analysis.

Finally, we found an almost significant ($p < 0.06$, after correction for multiple comparisons) ef-

[1]Note that the population is quite homogeneous, mean age=23.9 years and std=3.8 years. In this condition, the absence of an age effect at the group level is not surprising.

fect for the 3D rotation score in the activity in a sub-region the left occipito-parietal boundary [$z = 4.03, p < 3.10^{-5}$, uncorrected at $(-33, -65, 45)$mm]. Using the ROI-based analysis, the result was also present, and significant [$z = 3.10, p < 0.05$, corrected at $(-27, -73, 36)$mm], while this -and only this- ROI had a significant effect of the score on its position across subjects [$\chi^2_3 = 34.5, p < 0.05$, corrected at $(-27, -73, 36)$mm]. The regions that exhibit significant correlation of their signal or their position with the subject's sex and score in the 3D task are shown in Fig. 11 (a) and (b) respectively.

[Figure 11 about here.]

In summary, while voxel-based and ROI-based analyses show similar effect (in the statistical sense) of the age, sex or 3D task performance on the fMRI signal, the ROI-based analysis clearly indicates that these differences could be spatial, thus possibly anatomical, rather than merely quantitative.

VI. Discussion

Incrementing our knowledge on human brain function requires the common analysis of datasets from different subjects. While standard analyses assess the significance of effects at the voxel level, we show here that this procedure is not optimal since it suffers from the heterogeneity of the signal measured in different subjects, and from mis-registrations. To deal with these issues, taking into account the absence of a satisfactory generative model of brain activity in groups of subjects, we presented a rule-based, structural approach that extracts structures of interest in each subject's dataset and builds a group model from the structures of each subject.

Our solution is in the same spirit as a previous structural approach [30] based on the detection of scale-space blobs and the discovery of correspondences with a Markov Random Field (MRF). Our definition of activated regions by watershed analysis of supra-threshold regions is simpler and spatially better defined than the scale-space blobs. Our Belief Propagation scheme is quite comparable while simpler (see Fig. 3) than the MRF model [30], since the latter had to take into account some idiosyncrasies of the scale-space blob model. Furthermore our procedure inherits the good convergence properties of BP algorithms [36]. For the representation of the spatial structure of activated regions, we noticed that a loopy BP algorithm performed as well, or even slightly better than a tree-based BP algorithm; in any case, introducing the BP scheme markedly improved the reliability of the correspondences across groups with respect to a standard approach based only on the position in the common space (i.e. on Eq. (5)).

Other alternatives to our procedure are the cluster-based inference (CRFX, [18]) and parcel-based inference (PRFX, [7]). In particular, it is shown in Fig. 8 that these two are almost as reliable as the described approach. They suffer, however, from important drawbacks: cluster-based inference assumes that only wide supra-threshold clusters are worth reporting, which is not always true (see Fig. 6). Parcel-based inference that consists in making R/MFX tests on parcels instead of voxels, builds parcels of arbitrary size, and thus does not always correctly model the fine-scale activation pattern in each subject. The present approach based on watershed analysis of activity maps might reveal finer scale details (see Fig. 7). This might become especially important with the advent of high-resolution fMRI acquisition techniques [44]. Compared with the parcel-based approach for which the inter-subject correspondence is assumed a priori, the described solution relies on an a posteriori scheme that yields more information on the

reproducibility of a spatial pattern.

The specificity control of our procedure is correct on surrogate data (see Fig. 4), though much too conservative in many instances. This is due to simplifying (and conservative) assumptions used for the derivation of the test (see Sec. II-G). Finding a tighter upper bound of the error rate might be an important topic for future studies. Although first level statistics are used to define the regions of interest across subjects, an important point is that the control of false positive regions at the group level (see Eq. (9)) does not explicitly take the first-level statistics into account. In fact, the whole procedure, and in particular the statistical test described in section II-C is a region selection and association procedure that is blind to the first-level definition procedure (here, a p-value thresholding and watershed separation), and remains valid as along the first-level procedure is performed independently in each subject. In particular, it adapts quite automatically to the variability and the noise level in the dataset. Moreover, it should be stressed that first level statistics are usually not very reliable, because unmodeled effects (physiology/motion) can have a great impact on the effect significance estimation.

Our solution is well adapted to the characteristics of fMRI data, in the sense that it optimizes the compromise between sensitivity and specificity, as shown by the ROC curves in Fig. 5. The CR method outperforms the RFX thresholding procedure in most instances, especially when the homogeneity between subjects is low in terms of spatial or quantitative functional information. The only case where RFX outperforms the proposed method is in the absence of jitter, and for a weak control of specificity: in practice, both assumptions are unrealistic.

For the analysis of a standard group (10-15 subjects), see Fig. 6, this procedure is much more sensitive, and yields a much richer network than a more conventional approach (RFX, cluster-based RFX). The crucial point is that statistical tests (see Eq. 4) are performed on a reduced number of regions, allowing for a mild correction for multiple comparisons. It should also be noticed that the test is about the spatial density of local maxima of supra-threshold activity, and not the signal level or area of supra-threshold clusters: the spatial density of activated regions measures the reproducibility of an activation pattern across subjects, and seems to be a much more important feature than the average signal level across subjects. Moreover, such procedures that extract high-level features from the individual data and compare them across subjects are more robust to different pre-processing strategies, and/or to parameter tuning than traditional voxel-based methods.

The reliability of the detected areas in terms of reproducibility is higher than the reliability of voxel-based tests, as shown in Fig. 8. We obtained the same type of results for several other contrasts that involved motor, auditory or reading tasks and different numbers of subjects (not shown). An important feature of the region inference is thus that analyses performed in a group of $S = 10$ subjects should generalize to larger populations, while standard analyses show less reproducibility. This is crucial for the neuroimaging applications in both patients and normal subjects.

When used with a larger cohort of subjects, the CR method somewhat loses its advantage in sensitivity with respect to RFX methods, for two reasons: *i)* The test about activated regions is designed to select a certain *proportion* (α) of regions with high density of activity in the group, which limits the sensitivity of the method, while the regions selected by the RFX test will systematically increase with the number of subjects and asymptotically converge to all regions that have a possibly small, but positive effect (“half of the entire

brain") when the number of subjects increases; *ii)* finding stable configurations across the entire group of subjects is a much harder job, given the variety of the individual topographies; in particular, the replicator dynamics clique extraction procedure (see Sec. II-E) tends to oversegment the active regions, but this effect can be solved using hierarchical clustering instead. This problem might also be by-passed in the future e.g. by using multi-scale methods [30], [45]. It is important to note however that this effect is not significant for small sample sizes.

Our aim was also to refine the conventional point of view on the localization problem in fMRI data analysis [15]. In particular, an activity map of one subject in its native anatomical space is *not* comparable to a group map presented on an average anatomy: in the latter case, we present the locations where individual subjects drawn from the group *typically* activate, while in the former we present *for a specific subject* the regions with significant activity (see Fig. 7). This should help to better interpret fMRI group studies results. In standard analyses, it is impossible to distinguish between regions for which all subjects show a small increase of activity from regions for which only some subjects demonstrate increased activity.

Moreover, finding correspondences across subjects allows us to make statements on the dissimilarity between subjects [33], which is another *blind spot* of traditional M/RFX studies. Characterizing inter-subject differences in an interpretable way is essential if neuroimaging data is to be compared with genetic or behavioral information. An example is given in Sec. V-C, where one can make some inference on the between-subject variability by trying to explain differences in size/and or position of regions across subjects by some information that is available on these subjects. This kind of inference is possible, but quite cumbersome in the traditional voxel-based domain, due to a curse of dimensionality (the number of voxels is too high), and because it is not clear whether the variability can be attributed to differences in the signal level across subjects or to the position of the regions. Our procedure, for instance, indicated that the positions of the active regions, thus the functional anatomy, plays a non-negligible role in group discrimination. This point will be further studied in the future using e.g. multivariate classification/regression analysis techniques.

Another important question is whether this kind of analysis can be generalized to group comparison, which is important e.g. for the characterization of brain diseases. Although the answer is probably case-dependent, one possibility consists in pooling the subjects to define the ROIs, then derive subject or group *profiles* as in Sec. V-C, and to study possible group differences at the ROI level. However, we acknowledge that some cases may be problematic, e.g. if the inter-group differences reduce the sensitivity of the spatial test to discover some of the ROIs. On the other hand, the proposed method provides the opportunity to compare the regions position or shape across groups, which is not afforded by voxel-based models.

The present work is also an attempt to automatically find correspondences across subjects by associating activated areas with close relative positions across subjects. In particular, watershed analysis of the individual maps is used to define target regions. In order to enhance the understanding and interpretation of inter-subject variability, future developments might consider the use of generative models in the spatial and/or signal domain, based e.g. on Dirichlet Process Mixture Models of the fMRI data [27], [28]. Alternatively, a more anatomical point of view may be introduced, e.g. by defining the position of ROIs with respect to macro-anatomical features (sulco-gyral

anatomy [16]). This opens the way to an object-oriented representation of the functional anatomy [11]. It is also worthwhile to note that MNI/Talairach space does not play any particular role in the present method, so that any valid - and non-Euclidean - normalized space such as the one on the cortical surface [12], [46] can play the same role.

VII. Conclusion

In this work, we have shown that describing and comparing datasets with high level information instead of the usual voxel-based activity may benefit both the sensitivity and the reliability of fMRI group analyses. Moreover, this approach does not lose the spatial information through an averaging process, but enables neuroscientists to make explicit comparisons between subjects or groups of subjects in a more rigorous conceptual setting.

References

[1] J. Ashburner, K. Friston, and W. Penny, Eds., *Human Brain Function, 2nd Edition*. Academic press, 2004.

[2] K. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and A. Evans, "A general statistical analysis for fMRI data," *Neuroimage*, vol. 15, no. 1, pp. 1–15, January 2002.

[3] J. Ashburner and K. Friston, "Nonlinear spatial normalization using basis functions." *Hum Brain Mapp*, vol. 7, no. 4, pp. 254–66, 1999.

[4] J. Talairach and P. Tournoux, *Co–Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System : An Approach to Cerebral Imaging*. Thieme Medical Publishers, Inc., Georg Thieme Verlag, Stuttgart, New York, 1988.

[5] X. Wei, S.-S. Yoo, C. C. Dickey, K. H. Zou, C. R. G. Guttmann, and L. P. Panych, "Functional MRI of auditory verbal working memory: long-term reproducibility analysis." *Neuroimage*, vol. 21, no. 3, pp. 1000–8, Mar 2004.

[6] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, and J.-B. Poline, "Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses," *Neuroimage*, vol. 35, no. 1, pp. 105–120, 2007.

[7] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.-B. Poline, "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets," *Hum. Brain Mapp.*, vol. 27, no. 8, pp. 678–693, August 2006.

[8] P. Stiers, R. Peeters, L. Lagae, P. V. Hecke, and S. Sunaert, "Mapping multiple visual areas in the human brain with a short fMRI sequence." *Neuroimage*, vol. 29, no. 1, pp. 74–89, Jan 2006.

[9] D. L. Collins, L. G. G., and A. C. Evans, "Non-linear cerebral registration with sulcal constraints," in *MICCAI'98, LNCS-1496*, 1998, pp. 974–984.

[10] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher., D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson, "Retrospective evaluation of intersubject brain registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1120–1130, September 2003.

[11] J.-F. Mangin, D. Rivière, A. Cachia, E. Duchesnay, Y. Cointepas, D. Papadopoulos-Orfanos, D. L. Collins, A. C. Evans, and J. Régis, "Object-based morphometry of the cerebral cortex," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 968–982, August 2004.

[12] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis II: inflation, flattening and a surface-based coordinate system," *Neuroimage*, vol. 9, pp. 195–207, 1999.

[13] M. I. Miller, M. F. Beg, C. Ceritoglu, and C. Stark, "Increasing the power of functional maps of the medial temporal lobe by using large deformation diffeomorphic metric mapping." *Proc Natl Acad Sci U S A*, vol. 102, no. 27, pp. 9685–9690, Jul 2005.

[14] Y. Wang, X. Gu, K. M. Hayashi, T. F. Chan, P. M. Thompson, and S.-T. Yau, "Brain surface parameterization using riemann surface structure." *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv*, vol. 8, no. Pt 2, pp. 657–665, 2005.

[15] M. Brett, I. Johnsrude, and A. Owen, "The problem of functional localization in the human brain." *Nature Reviews Neuroscience*, vol. 3, no. 3, pp. 243–249, March 2002.

[16] D. Rivière, J.-F. Mangin, D. Papadopoulos-Orfanos, J.-M. Martinez, V. Frouin, and J. Régis, "Automatic recognition of cortical sulci of the human brain using a congregation of neural networks," *Medical Image Analysis*, vol. 6, no. 2, pp. 77–92, 2002.

[17] S. Mériaux, A. Roche, B. Thirion, and G. Dehaene-Lambertz, "Robust statistics for nonparametric group analysis in fMRI," in *Proc. 3th Proc. IEEE ISBI*, Arlington, VA, April 2006, pp. 936–939.

[18] S. Hayasaka and T. Nichols, "Validating Cluster Size Inference: Random Field and Permutation Methods," *Neuroimage*, vol. 20, no. 4, pp. 2343–2356, 2003.

[19] S. Mériaux, A. Roche, G. Dehaene-Lambertz, B. Thirion, and J.-

B. Poline, "Combined permutation test and mixed-effect model for group average analysis in fMRI." *Hum. Brain Mapp.*, vol. 27, no. 5, pp. 402–410, May 2006.

[20] R. Saxe, M. Brett, and N. Kanwisher, "Divide and conquer: A defense of functional localizers." *Neuroimage*, vol. 30, no. 4, pp. 1088–1096, May 2006.

[21] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain." *Neuroimage*, vol. 15, no. 1, pp. 273–89, Jan 2002.

[22] A. Nieto-Castanon, S. Ghosh, J. Tourville, and F. Guenther, "Region of interest based analysis of functional imaging data," *Neuroimage*, vol. 19, no. 4, pp. 1303–1316, 2003.

[23] B. Horwitz, K. Amunts, R. Bhattacharyya, D. Patkin, K. Jeffries, K. Zilles, and A. R. Braun, "Activation of broca's area during the production of spoken and signed language: a combined cytoarchitectonic mapping and pet analysis." *Neuropsychologia*, vol. 41, no. 14, pp. 1868–1876, 2003.

[24] K. Amunts, P. H. Weiss, H. Mohlberg, P. Pieperhoff, S. Eickhoff, J. M. Gurd, J. C. Marshall, N. J. Shah, G. R. Fink, and K. Zilles, "Analysis of neural mechanisms underlying verbal fluency in cytoarchitectonically defined stereotaxic space–the roles of brodmann areas 44 and 45." *Neuroimage*, vol. 22, no. 1, pp. 42–56, May 2004.

[25] S. B. Eickhoff, S. Heim, K. Zilles, and K. Amunts, "Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps." *Neuroimage*, vol. 32, no. 2, pp. 570–582, Aug 2006.

[26] W. Penny and K. Friston, "Mixtures of general linear models for functional neuroimaging." *IEEE Trans Med Imaging*, vol. 22, no. 4, pp. 504–514, Apr 2003.

[27] S. Kim, P. Smyth, and H. Stern, "A nonparametric bayesian approach to detecting spatial activation patterns in fmri data," in *Proc. 9th MICCAI*, ser. LNCS 4190. Copenhaguen: Springer Verlag, 2006, pp. 217–224.

[28] S. Kim and P. Smyth, "Hierarchical dirichlet processes with random effects," in *Advances in Neural Information Processing Systems*, Vancouver, 2006.

[29] O. Simon, F. Kherif, G. Flandin, J.-B. Poline, D. Rivière, J.-F. Mangin, D. L. Bihan, and S. Dehaene, "Automatized clustering and functional geometry of human parietofrontal networks for language, space, and number." *Neuroimage*, vol. 23, no. 3, pp. 1192–1202, 11 2004.

[30] O. Coulon, J.-F. Mangin, J.-B. Poline, M. Zilbovicius, D. Roumenov, Y. Samson, V. Frouin, and I. Bloch, "Structural group analysis of functional activation maps," *Neuroimage*, vol. 11, pp. 767–782, 2000.

[31] B. Thirion, P. Pinel, and J.-B. Poline, "Finding landmarks in the functional brain: Detection and use for group characterization," in *Proc. MICCAI2005*, Palm Spings, USA, October26-29 2005.

[32] L. Vincent, "Exact euclidean distance function by chain propagation," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 1991, pp. 520–525.

[33] F. Kherif, J.-B. Poline, S. Mériaux, H. Benali, G. Flandin, and M. Brett, "Group analysis in functional neuroimaging: selecting subjects using similarity measures," *Neuroimage*, vol. 20, no. 4, pp. 2197–2208, January 2004.

[34] M. Woolrich, B. Ripley, M. Brady, and S. Smith, "Temporal autocorrelation in univariate linear modelling of fMRI data," *Neuroimage*, vol. 14, no. 6, pp. 1370–1386, December 2001.

[35] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, Inc., 1988.

[36] J. Yedidia, Y. Weiss, and W. T. Freeman, "Understanding belief propagation and its generalizations," in *International Joint Conference on Artificial Intelligence*, 2001.

[37] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 1, pp. 167–172, Jan. 2007.

[38] G. Lohmann and S. Bohn, "Using replicator dynamics for analyzing fMRI data of the human brain." *IEEE Trans Med Imaging*, vol. 21, no. 5, pp. 485–492, May 2002.

[39] J. Neumann, G. Lohmann, J. Derrfuss, and D. Y. von Cramon, "Meta-analysis of functional imaging data using replicator dynamics." *Hum Brain Mapp*, vol. 25, no. 1, pp. 165–173, May 2005.

[40] M. Pelillo and A. Torsello, "Payoff-monotonic game dynamics and the maximum clique problem," *Neural Computation*, vol. 18, no. 5, pp. 1215–58, May 2006.

[41] M. Liou, H.-R. Su, J.-D. Lee, P. E. Cheng, H. C.-C., and C.-H. Tsai, "Bridging functional MR images and scientific inference: Reproducibility maps," *Journal of Cognitive Neuroscience*, vol. 15, no. 7, pp. 935–945, 2003.

[42] C. R. Genovese, D. C. Noll, and W. F. Eddy, "Estimating test-retest reliability in functional MR imaging. I: Statistical methodology." *Magn Reson Med*, vol. 38, no. 3, pp. 497–507, September 1997.

[43] C. Lemer, S. Dehaene, E. Spelke, and L. Cohen, "Approximate quantities and exact number words: dissociable systems." *Neuropsychologia*, vol. 41, no. 14, pp. 1942–1958, 2003.

[44] K. Grill-Spector, R. Sayres, and D. Ress, "High-resolution imaging reveals highly selective nonface clusters in the fusiform face area." *Nat Neurosci*, vol. 9, no. 9, pp. 1177–1185, Sep 2006.

[45] J.-B. Poline and B. M. Mazoyer, "Analysis of individual brain activation maps using hierarchical description and multiscale detection," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 702–710, December 1994.

[46] C. Clouchoux, O. Coulon, D. Rivière, A. Cachia, , J.-F. Mangin, and J. Régis, "Anatomically constrained surface parameterization for cortical localization," in *MICCAI'05*, 2005, pp. 344–351.

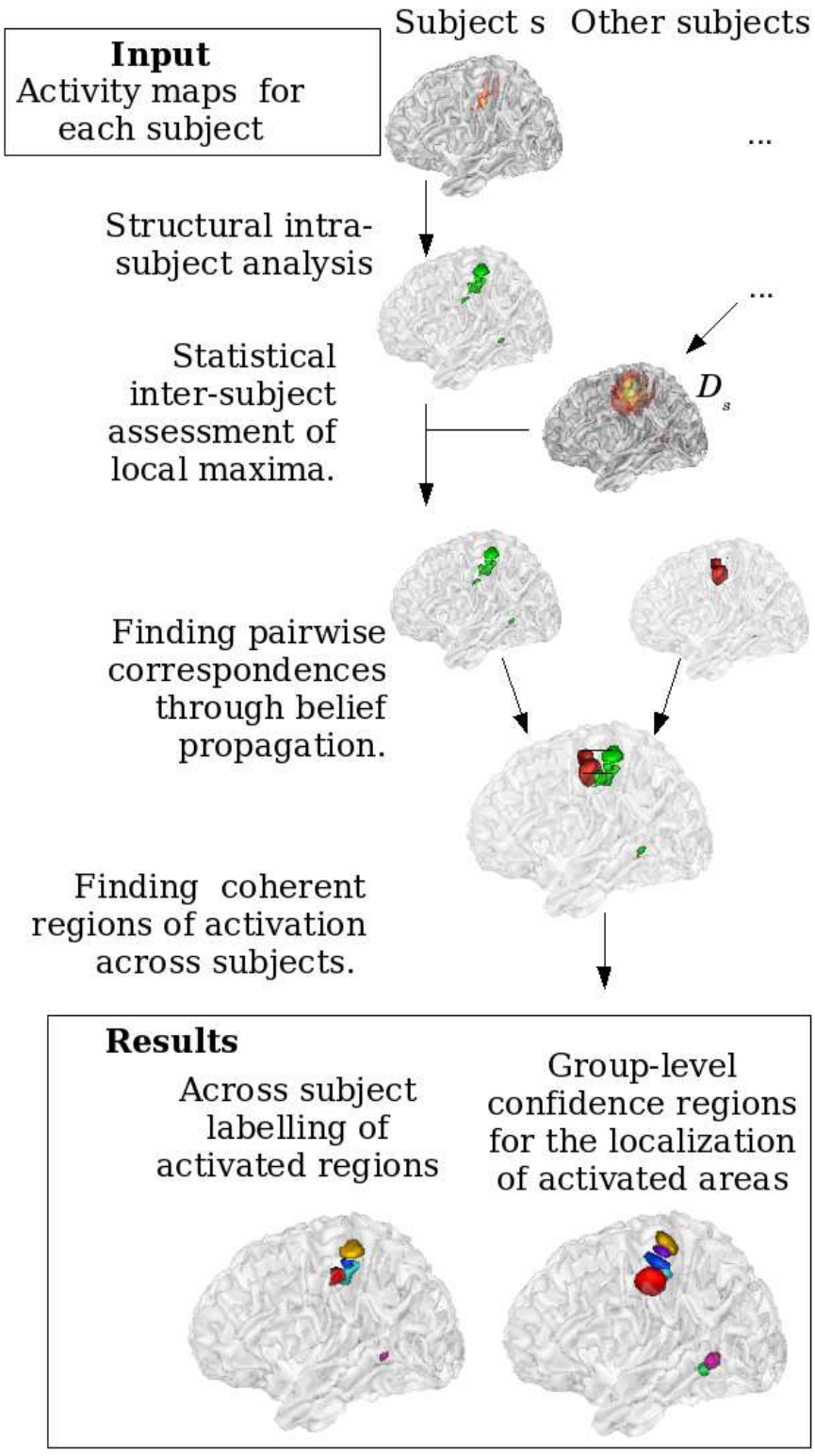


Fig. 1. Flowchart of our method for structural analysis of group data. This is a pipeline, or set of procedures, that produces a group-level representation of individual activation patterns based on reproducibility analysis; we illustrate it for a motor activation study. The input to the method consists of activation images, one for each subject. An intra-subject structural analysis is first performed, resulting in a set of activated regions. The cross-subject spatial density of activated regions is derived, and only the maxima that fall in the highest density regions are further considered. Probabilistic correspondences are then found between the regions of each pair of subjects, using a belief propagation algorithm. Finally, associated regions are segregated into inter-subjects cliques, so that each region of reproducible activity is labeled consistently across subjects, and confidence regions for the position of these ROIs are derived at the group level.

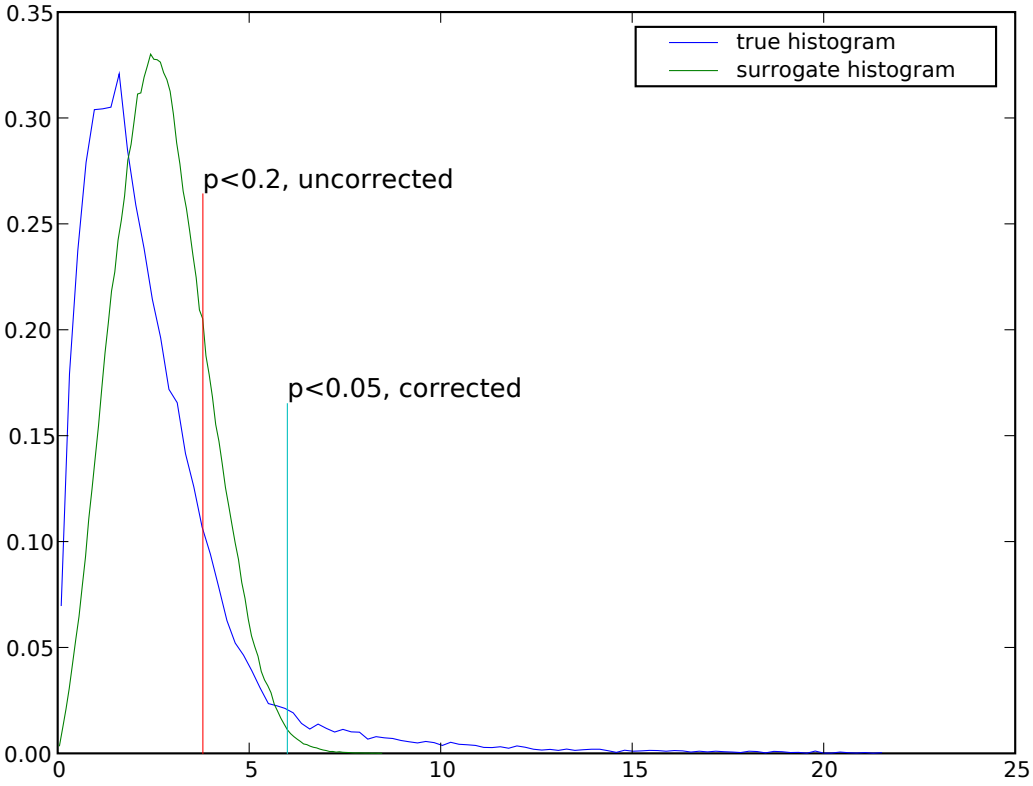
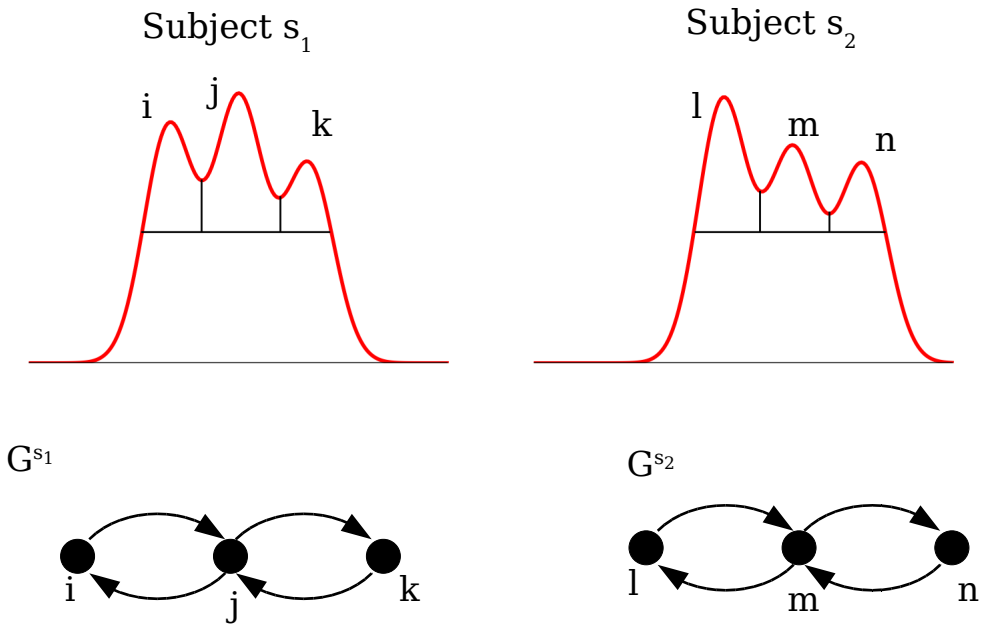


Fig. 2. Modeling the density of $\mathcal{D}_s$ in the volume. Randomly reshuffling the position of the maxima of activity in subjects $\sigma \in \{1, .., S\} - \{s\}$ yields an empirical histogram of $\mathcal{D}_s$ under the null hypothesis (green), which can be used to define critical values u_α, which can be corrected for multiple comparisons or not, to threshold the density $\mathcal{D}$.



Initialization

	$P(a_l^{s_2} \leftarrow .)$	$P(a_m^{s_2} \leftarrow .)$	$P(a_n^{s_2} \leftarrow .)$
$P(. \leftarrow a_i^{s_1})$	**0.59**	0.31	0.10
$P(. \leftarrow a_j^{s_1})$	**0.42**	0.38	0.20
$P(. \leftarrow a_k^{s_1})$	0.25	**0.39**	0.35

After convergence

	$P(a_l^{s_2} \leftarrow .)$	$P(a_m^{s_2} \leftarrow .)$	$P(a_n^{s_2} \leftarrow .)$
$P(. \leftarrow a_i^{s_1})$	**0.74**	0.23	0.03
$P(. \leftarrow a_j^{s_1})$	0.41	**0.46**	0.13
$P(. \leftarrow a_k^{s_1})$	0.14	0.40	**0.46**

Fig. 3. Illustration of the use of the Belief Propagation algorithm to find correspondences between maxima within a pair of subjects. This is a toy dataset, in a one-dimensional space. The activity maps of subjects s_1 and s_2 are shown on the top of the figure, together with a watershed segmentation. In that case, $t_i^{s_1} = 0, t_j^{s_1} = 1$ and $t_k^{s_1} = 2$, while $t_l^{s_2} = 0.7, t_m^{s_2} = 1.7$ and $t_n^{s_2} = 2.7$; $\delta_\tau = 1.4$. The related graphs G^{s_1} and G^{s_2} are shown below; in this case they are isomorphic. The associations are initialized using Eq. (5), and then refined using Eq. (6): Clearly the message passing algorithm enhances the probabilities $P(a_l^{s_2} \leftarrow a_i^{s_1}), P(a_m^{s_2} \leftarrow a_j^{s_1})$, and $P(a_n^{s_2} \leftarrow a_l^{s_1})$, thus compensates the effect of the global translation between the two datasets. This effect is also present with the converse probabilities $P(a_\cdot^{s_1} \leftarrow a_\cdot^{s_2})$.

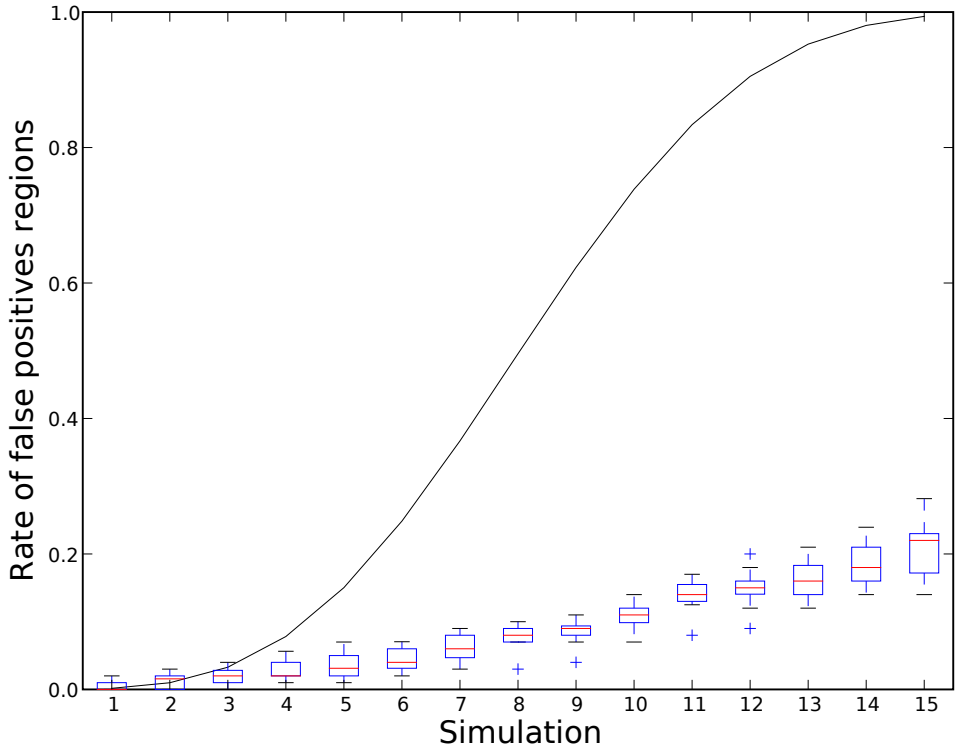
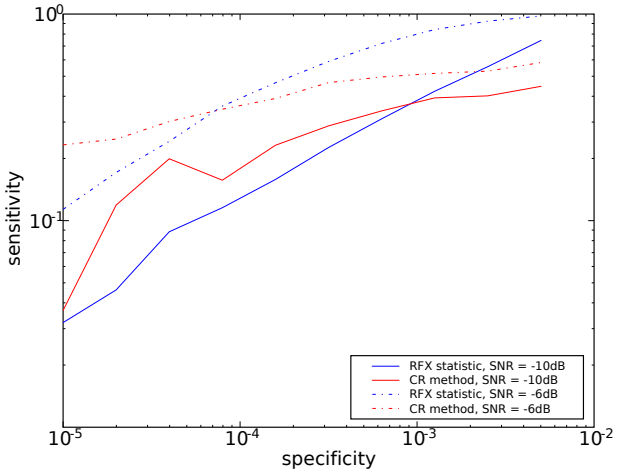
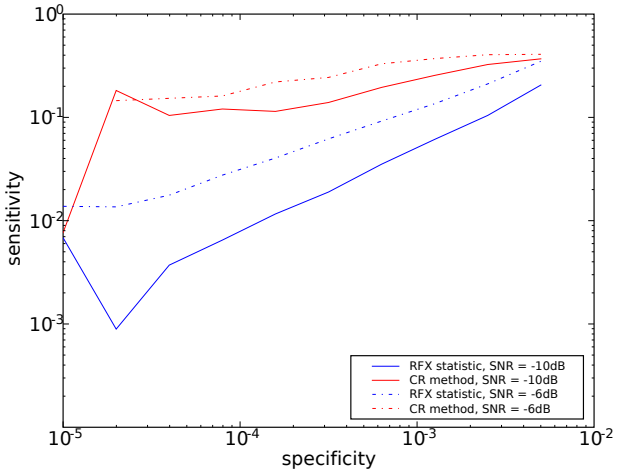


Fig. 4. Number of false detections obtained with our method in noise only environments (box and whisker plot), compared with its expected value (continuous line), for values of α ranging from 0.1 to 0.8 in 0.05 steps. This is based on 20 runs of 100 simulations under the null hypothesis. The fact that the box plots lie beneath the line shows that the threshold is rather conservative, especially for large values ($p > 0.2$).

(a)



(b)



(c)

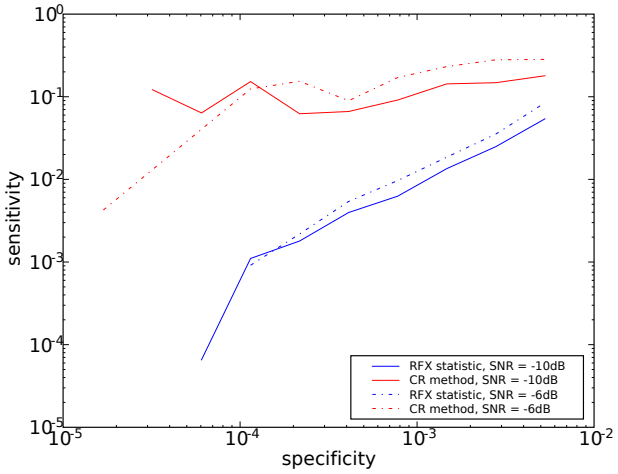
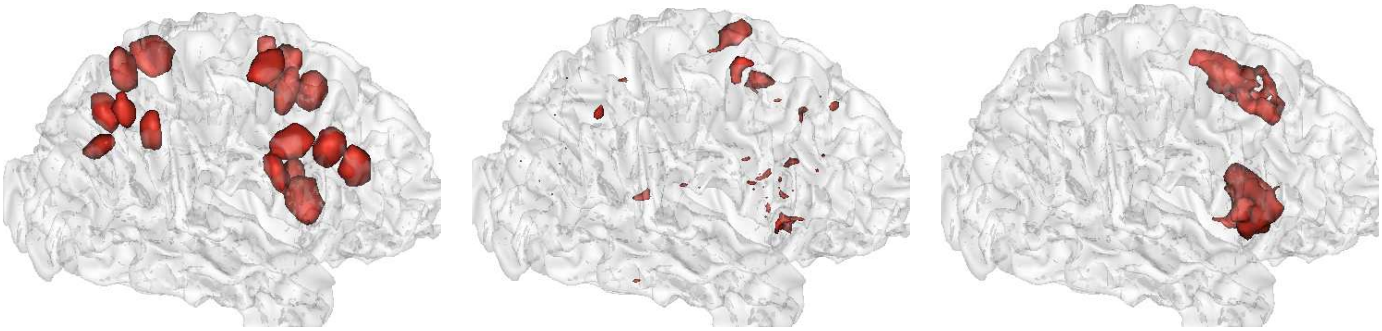
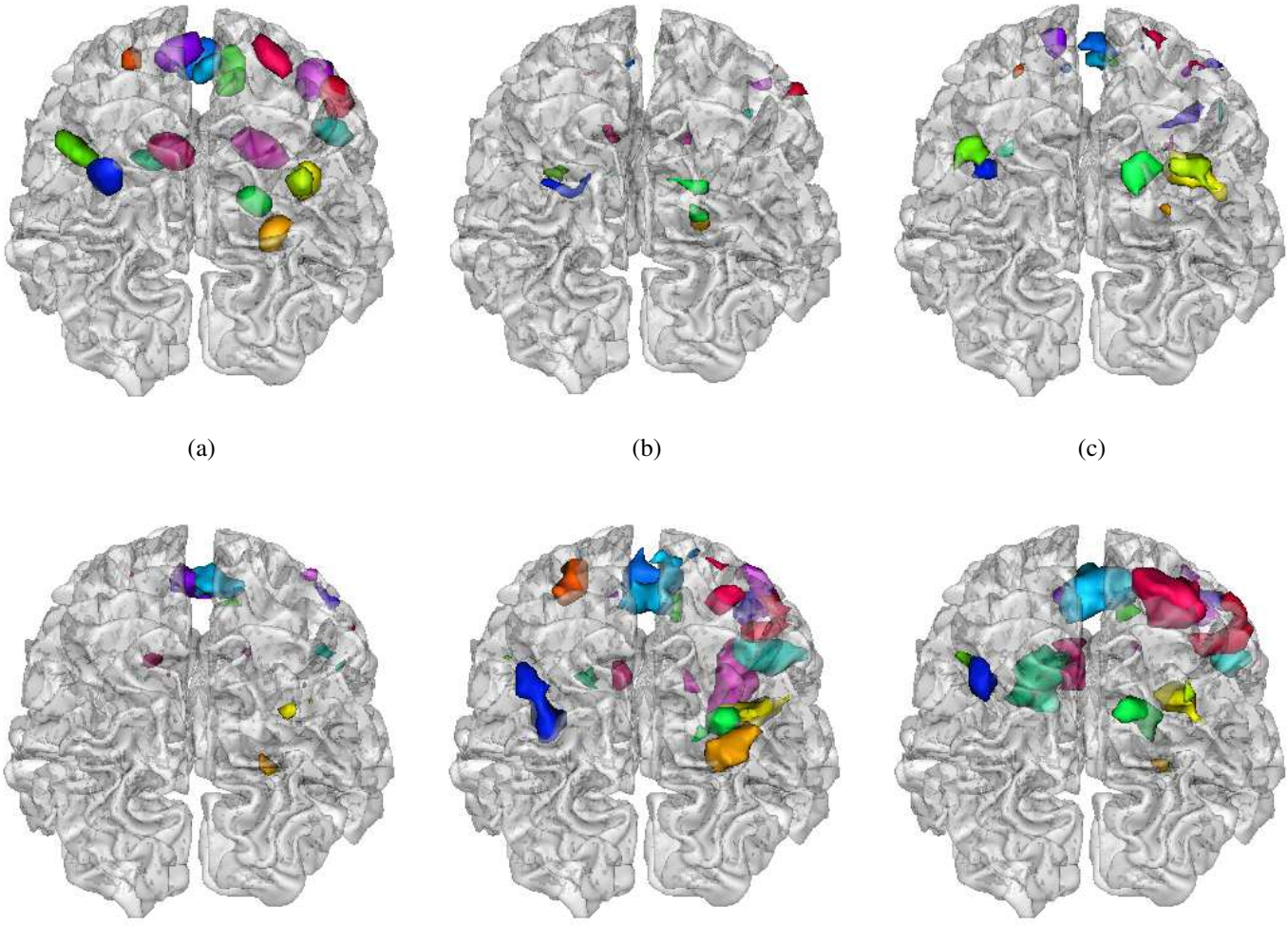


Fig. 5. ROC curves for the RFX and CR maps for different SNR levels and across-subject jitter magnitudes. (a) with no jitter of the activation position, ROC curves are presented for the RFX (blue) and CR method (red), for a mean SNR of -6dB (continuous line) or -10dB (dashed line). (b) and (c): the same curves, with a jitter of magnitude of 1 voxel in each direction (b), or 2 voxels in each direction (c). Except for the situation with no jitter, the CR method outperforms the RFX method. In the case of no jitter, the CR method performs better at high specificity levels, but plateaus at a lower level when the specificity control is weaker.



(a) (b) (c)

Fig. 6. Comparison of the a CR map and RFX maps obtained for a functional contrast that shows regions involved in computation task. This is based on 10 subjects, with one session per subject. (a) The CR map, significant at $p < 0.05$, corrected level, shows $q = 19$ active regions; (b) the RFX map is thresholded at the voxel level at $p < 10^{-3}$ level, uncorrected (at $p < 0.05$ corrected, the map is empty); (c) the RFX map is thresholded at the cluster level at $p < 0.05$ level, corrected. The CR map clearly shows symmetric parietal regions involved in the computation task, while these regions are not detected with the RFX procedure.



(a) (b) (c)

(d) (e) (f)

Fig. 7. Active regions found at the group level and in five subjects of the dataset. (a) At the group level, 19 regions are spatially defined by their confidence ellipsoids. (b-e) This corresponds to regions that are present or not in each subject's dataset. Corresponding regions have the same color. Note that, besides differences in size and precise position, the relative positions are well preserved.

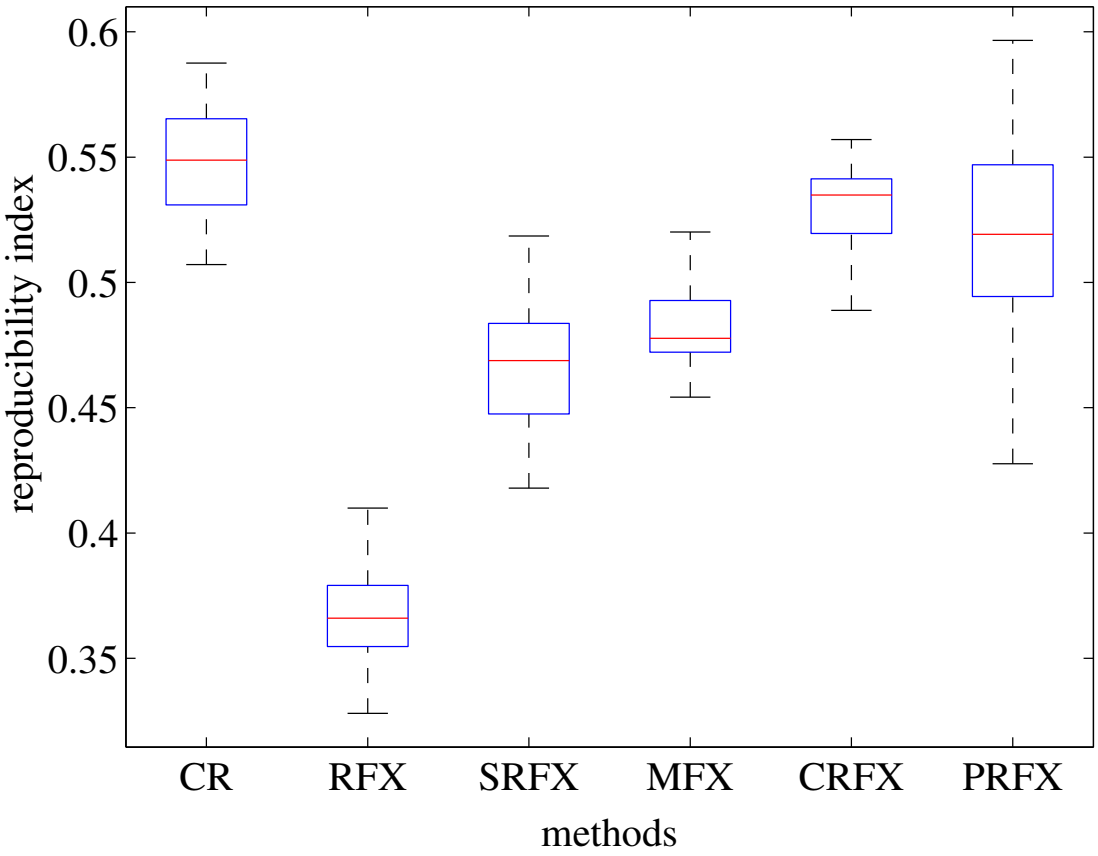
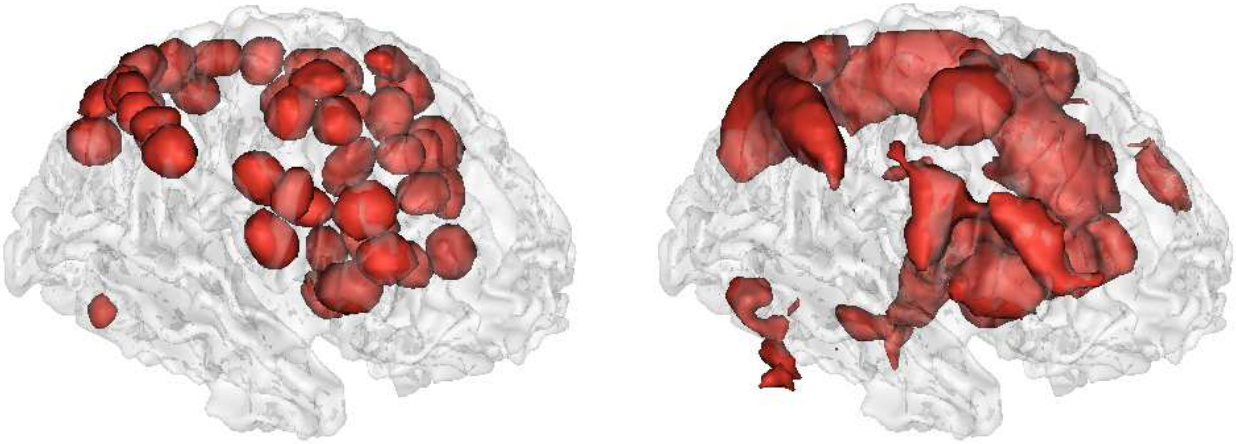


Fig. 8. Reproducibility indexes obtained by jackknife subsampling analysis of the population of 102 subjects in groups of 10 subjects, for six different techniques: our new technique based on confidence regions (CR), Random Effects Analysis (RFX), RFX after 12mm smoothing (SRFX), Mixed Effects Analysis (MFX), RFX analysis with cluster-level thresholding (CRFX) and Parcel-based RFX (PRFX).



(a) (b)

Fig. 9. Results of the group analysis that shows regions activated for a computation task across 102 subjects. (a) Confidence regions obtained with our approach, at $p < 0.05$; (b) Supra-threshold regions of the RFX map for this group of subjects, thresholded at the $p < 0.05$, corrected, voxel-level. The images are superimposed on a typical grey-white matter interface.

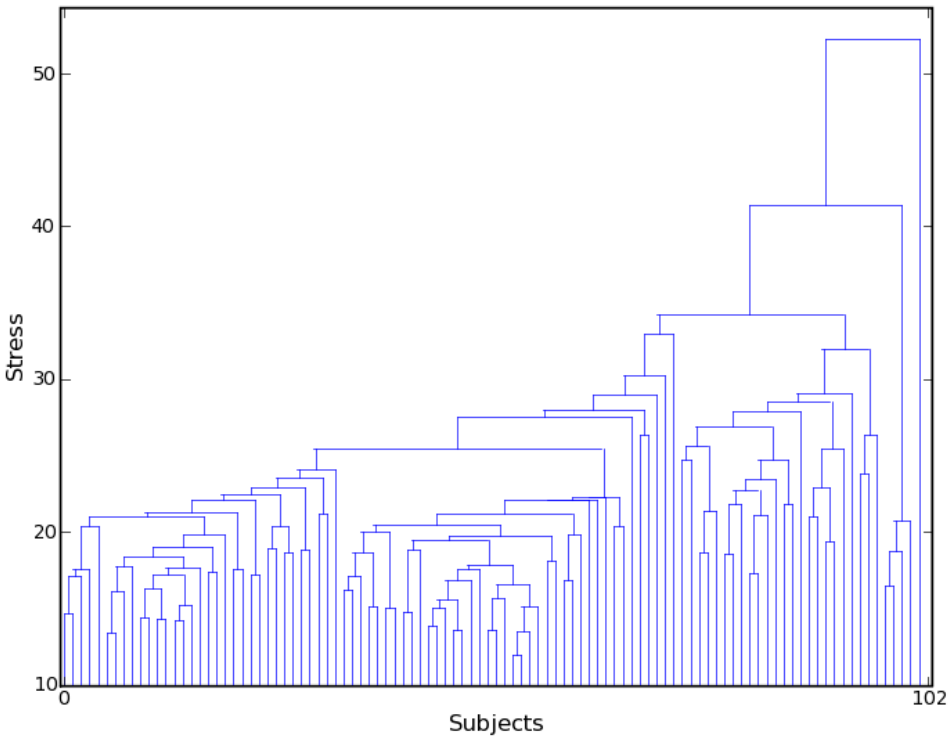
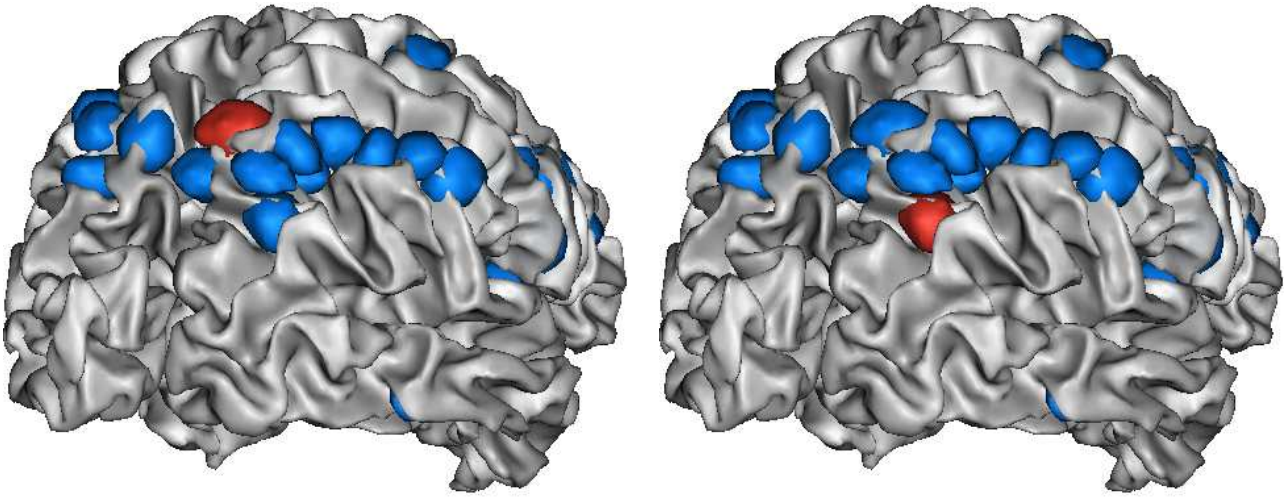


Fig. 10. Unsupervised classification of the group of subjects, based on their *profile*. The dendrogram shows the organization of the population in terms of hierarchical clustering, based on an average linkage approach. The dendrogram shows that there is one main group, plus a few scattered subjects (on the right side).



(a) (b)

Fig. 11. ROI whose activity and position is significantly modulated by the sex of the subjects (a) or their ability to perform a 3D task (b). Voxel-based analyses yield similar regions, but ROI-based increases the significance in case (b) and enables us to study the effect of regressors of interest on the position of the ROIs.

MNI coordinates	Anatomical location	hemisphere	number of subjects (/102)
(-54, -55, -10)	Temporal Inf	R	40
(52, 9,15)	Frontal Inf Oper	L	51
(-42, 35, 16)	Frontal Inf Tri	R	78
(-41, 27, 30)		R	71
(-29, -4, 54)	Frontal Mid	R	74
(-28, 10, 53)		R	51
(30, 4, 52)		L	64
(20, 3, 62)	Frontal Sup	L	39
(-21, 0, 63)		R	57
(-55, 2, 20)	Precentral	R	71
(-51, -4, 38)		R	70
(-51, 7, 34)		R	44
(-41, 6, 28)		R	49
(46, 7,31)		L	69
(42, -50, 51)	Parietal Inf	L	54
(51, -37, 48)		L	74
(31, -67, 41)		L	51
(42, -40, 40)		L	68
(34, -61, 51)		L	48
(-49, -42, 49)		R	59
(-42, -50, 50)		R	55
(-35, -47, 40)		R	72
(-35, -57, 55)	Parietal Sup	R	63
(-26, -63, 53)		R	40
(-23, -72, 48)		R	51
(15, -68, 52)		L	56
(33, 21, 1)	Insula	L	66
(-33, 19, 3)		R	70
(17, 14, 0)	Putamen	L	61
(-22, 5, -3)		R	45
(-5, -70, 45)	Precuneus	R	46
(-10, -66, 56)		R	63
(-16, 13, 0)	Caudate	R	48
(-13, 1, 8)		R	44
(12, 3, 9)		L	46
(-27, -73, 36)	Occipital Sup	R	77
(0, 1, 32)	Cingulum Mid	L	42
(6, 22, 38)		L	48
(-5, 21,40)		R	42
(-3, 28, 28)		R	45
(-55, -34, 43)	SupraMarginal	R	71
(4, 12, 50)	Supp Motor Area	L	60
(-5, 1, 53)		R	42
(-2, -1, 62)		R	51
(-23, -58, 45)	Angular	R	42

TABLE I

SUMMARY OF THE CONFIDENCE REGIONS FOUND FOR THE COMPUTATION TASK IN THE POPULATION OF $n = 102$ SUBJECTS. FOR EACH REGIONS, WE GIVE AN ANATOMICAL DESIGNATION, THE AVERAGE POSITION IN THE MNI COORDINATE SYSTEM, AND THE NUMBER OF SUBJECTS IN WHICH THIS REGION CAN B FOUND. R STANDS FOR RIGHT, L STANDS FOR LEFT, INF. FOR INFERIOR, MID. FOR MIDDLE, SUP. FOR SUPERIOR, SUPP. FOR SUPPLEMENTARY, OPER FOR OPERCULUM, TRI FOR TRIANGULAR.